International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 03 - Issue 01 || January 2018 || PP.05-12

An Efficient Sentimental Analysis Using Dynamic Cluster Based Naive Bayesian Classification

S.Indhu
M.Phil Research Scholar,
Department of Computer Science
Sri Ramakrishna College of Arts and Science for women
Coimbatore, India

Mrs.S.R.Lavanya
Assistant Professor,
Department of Computer Science
Sri Ramakrishna College of Arts and Science for women
Coimbatore, India

Abstract: Social media is one of the most important forums to convey opinions. Sentiment analysis is the procedure by which information is obtained from the opinions, assessment,and feelings of individuals in regards to entities, events and their features. Analyzing sentimental words is to examine and cluster the user created information like reviews, blogs, comments, articles etc. This research work presents an optimal Dynamic Cluster based Naive Bayesian (DCNB) classification model to deal with the troubles in one go under a combined framework. DCNB represents each review document in the form of opinion pairs, and can concurrently model aspect terms and corresponding opinion words of the review for hidden aspect and sentiment detection.Meanwhile, the proposed system processed meaningful tweets into three different clusters positive, neutral and negative using unsupervised machine learning technique such as clustering.
KEYWORDS: Social media, Twitter, Sentiment Analysis, BayesianClassification.

I. INTRODUCTION

Microblog services like twitter invite more persons to post their emotions and sentiments on different topics. The posting of sentiment contents can not only give an emotional snapshot and also have potential commercial [1], financial [3] and sociological values [4]. But facinga large number of sentiment tweets is hard for individuals to get anoverall impression. As a result, there are many sentiment classification works showing interests in tweets [5].

Twitter allows different types of topics to discuss. Sentiment classifiers always devote themselves to a particular domain or topic. Specifically, sentiment data from one topic performs ineffectively on test data from another. Because words and languages construct used for expressing sentiments can be unique on various themes. For example, “Samsung is good” it could be confident in a product survey but negative in a movie survey. In social media, a Twitter user may have different opinions on different topics [9]. Hence, thetopic adjustment is required for sentiment classification of tweets on developing and unpredictable topics.

Generally, sentiments and opinions can be evaluated at various levels of granularity. The assignment of evaluating overall sentiments of texts is usually defined as aclassification problem, e.g., classifying a review document into positive or negative sentiment. Then, a variety of machine learning methods trained using different types of indicators (features) have been employed for overall sentiment analysis [2], [6], [7] and [8].

In recent years sentiment analysis has more interest in academic and industry because of its potential applications. In figure 1 shows the structure of sentiment classification framework. One of the most promising applications is ananalysis of opinions in social networks. Lots of individuals write their feelings in forums, microblogging or review websites.

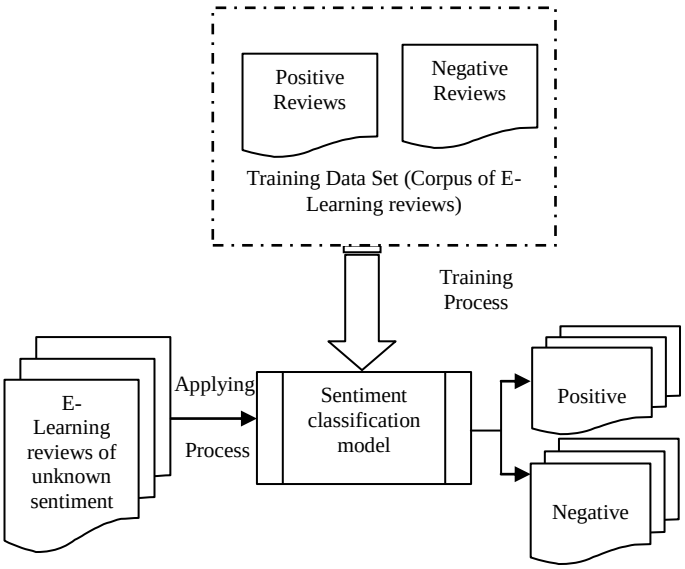


Figure 1: Sentiment classification Framework

This information is very helpful for organizations, governments, and people, who need to track mechanically attitudes and feelings in those locations. To be specific, there is a lot of information accessible that covers many helpful data, so it can be examined mechanically. For example, a customer who needs to purchase an item generally looks through the Web trying to discover opinions of other customers or reviewers about this item. In fact, these sorts of reviews affect customer's choice.

Sentiment mining is a computational study of opinions, sentiments,and emotions expressed in texts. Feelings exist on the Web for any person or product and for the features or modules of these products similar to a cell phone battery, touchscreen display, and keyboard, etc. Identifying sentiments is viewed as a tough task. For example, 'la application responde muy rápido (the application reacts quickly)'; the sentiment of the opinion is confident because the word ‘rápido (quick)’ suggests a good thing- it is good that applications run quick. But, a similar word in another context, as in the sentence 'la bateríase descargómuyrápido (the battery released quickly)', infers a negative assumption—unfortunately that batteries decrease their energy rapidly. Thus, the issue suggests utilizing of world knowledge is exceptionally huge and complex issue.

This paper presents an optimal Dynamic Cluster based Naive Bayesian (DCNB) classification method.It is applied to analysis unsupervised sentiment classification.

The main contributions of this paper are as follows:

- This work presents a novel Dynamic Cluster based Naive Bayesian (DCNB) classification called *DCNB*, which forms the calculation for overall ratings/sentiments of reviews via non- linear model based on the conditional hidden aspects and sentiments in the reviews.
- This paper proposes feature extraction process with text cleaning strategy, which can be used to represent and classifyuncomplicated and interpretable noisy data.

II. RELATED WORK

In [2] authors considered the issue of classifying documents not by topic, but by overall feelings. For instance, deciding whether a review is positive or negative. Utilizing movie reviews as information, they discover that standard machine learning systems conclusively beat human-produced baselines. However, the three machine learning methods employed here (Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization.

In [3] authors explained the sparsity issue by abusing term relationships along with Naive Bayes classifiers. The primary strategy is to appraise term relationships based on the co-occurrence information of two terms in a specific context. The second technique assesses the term relationships based on the distribution of terms over various hierarchical categories in publicly accessible document taxonomy. Then term relationship is utilized to increase Naive Bayes classifiers. To test their methods on two open-domain data sets to exhibit its points of interest.

In [4] an author explained about Twitter is a microblogging site where people read and compose lots of short messages on a collection of topics every day. This experiment utilizes the circumstance of the German federal election to research whether Twitter is used as a forum for political pondering and whether online messages on Twitter truly reflect offline political opinion. With LIWC text analysis software, they conducted a content-analysis of more than 100,000 messages covering a reference to either a political party or politician. The outcomes demonstrate that Twitter is indeed used extensively for political thought. And found that the mere number of messages specifying a party imitates the election result.

In [5] authors introduced a methodology of building statistical models from the social media dynamics to forecast collective sentiment dynamics. This paper model the collective sentiment change without delving into microanalysis of single tweets or clients and their corresponding low-level system structures. Experiments on large-scale Twitter data demonstrate that the model can accomplish above 85% precision on directional sentiment estimate.

In [6] authors introduced a model that uses both the supervised and unsupervised methods to learn word vectors capturing semantic term—document information as well as rich feeling content. The proposed model can use both continuous and multi-dimensional sentiment information and also non-sentiment annotations. To instantiate the model to use the document-level sentiment polarity annotations introduced in several online documents (e.g. star ratings). To assess the model utilizing little, widely used sentiment and subjectivity corpora and discover it overtakes several previously presented strategies for sentiment classification. It also presents an enormous dataset of movie reviews to serve as a more robust benchmark for work in this area.

In [7] authors presented a context-aware method for analyzing sentiment at the level of individual sentences. Most existing machine learning methods suffer from limitations in the modeling of complex linguistic structures across sentences and often fail to capture non-local contextual cues that are significant for sentiment interpretation. In contrast, this approach permits structured modeling of emotion while considering both local and global contextual information. In particular, it encodes intuitive lexical and discourse knowledge as expressive constraints and incorporates them into the learning of conditional random field models via posterior regularization. The context-aware constraints provide additional power to the CRF model and can manage semi-supervised learning when labeled data is restricted.

In [8] authors defined a probabilistic model Latent Dirichlet Allocation (LDA), is a model for grouping of separate data, for instance, text corpora. LDA is a three-level hierarchical Bayesian model, in which each component of a group is revealed as a limited mixture over an underlying set of topic probabilities. Each topic is shown as an immeasurable mixture over an underlying set of topic probabilities. In the circumstance of text modeling, the topic probabilities provide an obvious depiction of a text. To present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation.

In [10] authors focused on thereproduction of user-generated evaluation and generally rating pairs, and intend to recognize semantic features and aspect-level sentiments from review data as well as to calculate overall sentiments of reviews. They proposed a new probabilistic supervised joint aspect and sentiment model (SJASM) to compact with the problems in one go below a unified framework. SJASM signifies every review document in the appearance of opinion pairs, and can concurrently model feature terms and equivalent opinion words of the evaluation for hidden aspect and sentiment detection. It also leverages sentimental overall ratings, which frequently come with online reviews, as direction data, and can gather the semantic aspects and aspect-level sentiments that are not only significant but also analytical of overall sentiments of reviews. Furthermore, they also developed proficient inference technique for parameter evaluation of SJASM based on collapsed Gibbs sampling.

In [11] Neural network methods have achieved proficient results for sentiment classification of text. Though, these representations only employ semantics of texts, while overlooks users who communicate the sentiment and products which are assessed, both of which have large control on understanding the sentiment of text. In this paper, authors addressed these concerns by integrates user and product level information into a neural network approach for document-level sentiment classification. Consumers and products are models using vector space models, the demonstrations of which capture significant global clues such as personality preferences of users or general qualities of products. Such global confirmation, in turn, facilitates embedding knowledge procedure at document rank, yielding improved text representations. By joining confirmation at theuser, product and document level in an integrated neural framework, the proposed model attains state-of-the-art performances on IMDB and Yelp datasets.

In [12] Internet is regularly used as an average for substitute of information and opinions, as well as propaganda propagation. In this learning, the use of sentiment analysis methodologies is proposed for categorization of Web forum opinions in several languages. The effectiveness of stylistic and syntactic attributes is assessed for sentiment classification of Arabic and English content. Exact feature extraction mechanisms are included to anexplanation for the linguistic characteristics of Arabic. The entropy weighted genetic algorithm

(EWGA) is also enlarged, which is hybridized genetic algorithms that integrate the information-gain heuristic for attribute selection. EWGA is planned to progress performance and obtain a better evaluation of key features. The proposed features and methods are assessed on a standard movie review dataset and theU.S. and Middle Eastern Web forum postings.

In [13] authors presented a novel approach to phrase-level sentiment study that first decides whether an expression is unbiased or polar and then disambiguates the division of the polar terms. With this technique, the scheme is proficient to mechanically recognize the related polarity for a huge subset of sentiment expressions, attaining results that are extensively improved than baseline.

In [14] authors presented a classifier to calculate related polarity of subjective expressions in a sentence. This approach attributes lexical scoring resulting from the Dictionary of Affect in Language (DAL) and complete throughout Word Net, permitting us to repeatedly achieve the huge common of words in the input circumvent the requirement for manual labeling. They enhance lexical rating with n-gram analysis to discover the result of environment. To join DAL scores with syntactic elements and then extract n-grams of elements from all sentences. It also employs the polarity of all syntactic constituents within the sentence as attributes.

In [15] sentiment analysis of product reviews, one significant problem is to create a review of opinions based on product features (also called characteristics). Nevertheless, for the similar feature, thepublic can convey it with several different words or phrases. To create a useful outline, these words and phrases, which are domain synonyms, requiresbeing grouped under the similar feature group. Though numerous techniques have been proposed to extract product features from reviews, incomplete work has been done on clustering of synonym attributes. This paper focused on this task. Typical methods for solving this difficulty are based on unsupervised learning using various forms of distributional correspondence. Nevertheless, they establish that these techniques do not do well. This representation is a semi-supervised learning problem. Lexical characteristics of the difficulty are exploited to without human intervention identify some labeled examples.

III. RESEARCH METHODOLOGY

This research work presentsan optimal Cluster based Bayesian Classification method which integratesinto the sentiment analysis classification. It forms the calculation for overall ratings/sentiments of reviews via non- linear model based on the conditional hidden aspects and sentiments in the reviews. This method follows a well-establishedNaive Bayesian classification with annotationprocess makes thelevel of underlying linguistic representation of text.The proposed architecture diagram is described in figure 2.

A. Twitter Data Extraction Process

Twitter data extraction is to get the tweets using Twitter API to collect the information from various hash-tags like movie reviews. In this process, the twitter raw data's are extracted from twitter website. This dataset contains many sentences labeled with 0 or 1 or -1 depending on its polarity, and since the sentences were shorter than 160 characters and they were extracted from social media and used them as tweets.

Although the main purpose of this extraction process was mostly to build the Twitterweb server. The experiment with Twitter API is to pull specific tweets in real-time for further classification using an external sentiment analysis classification tool. The dataset consists of more than 100 tweets with the hashtag, which were used for sentiment analysis classification. This dataset was unique, and every single tweet was organic, there were some retweets and so on, but in general, it was really a clean collection of tweets. However, the worst was that most of the tweets were positive and that was not good to train the model, the ideal is to have a balance between the different classes.

In order to have right to use the twitter data programmatically, this requiresgenerating an application that cooperates with the Twitter API. To Sign in to twitter account it will provide the consumer key and consumer secret key described below these are the submission settings that should be reserved privately. From the configuration page of an application, it can also precede a permit token and an access token secret. Likewise, the consumer keys strings must also be kept as private: they present the application access to Twitter on behalf of an account.

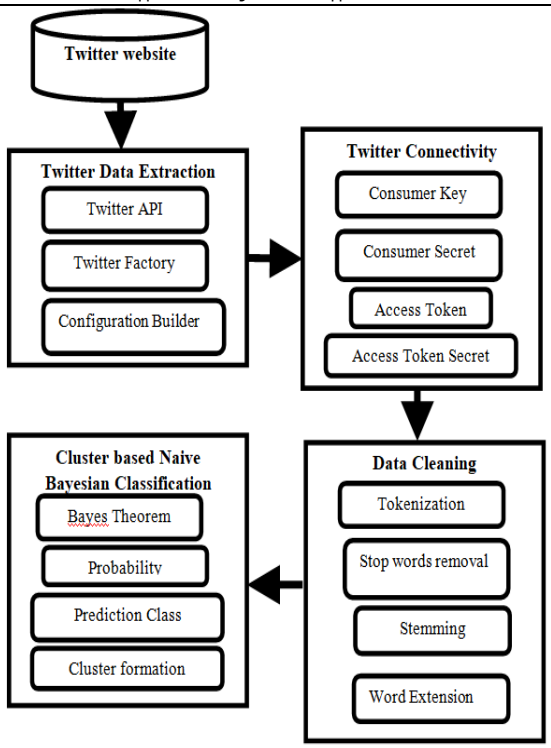


Figure 2: Proposed Architecture diagram

```
ConfigurationBuilder cb = new ConfigurationBuilder();
cb.setOAuthConsumerKey("5J52DKViHzPXavEIYGtv8DaYv");
cb.setOAuthConsumerSecret("YoJ7FNR9Eaf2LHhHAlLP93I7Li
QN1l2Gcy65wsrYiXPMnDx98B");
cb.setOAuthAccessToken("2926600375-
dtjAUXoCvq66YNJ7p4Hqmu3YU8y93fCKQPu0ESQ");
cb.setOAuthAccessTokenSecret("NCH6BscKVySfvIDTWpUCx
WIQMWyYm8vuOIBMCeqBChu8z");
```

Configure secret keys

B. Data Cleaning Process

Data cleaning or preprocessing is the most important part of this paper. Preprocessing of data is the process of planning and cleaning the tweets for clustering. Reducing the irrelevant noise in the tweets should help to progress the performance of the clustering and speed up the clustering process. This process performed the following processes on tweets during cleaning and normalization.

Tokenization: The given input as tweet character progression, tokenization is anassignment of splitting it up into parts called tokens and at the same momentclearing certain characters such as punctuation marks.

Stop Words Removal: A stop-list is the name frequently given to a position or list of stop words. It is naturally language specific, although it may contain words. Some of the various commonly used stop words from English include “a”, “of”, “the”, “I”, “it”, “you”, “and”, these are generally regarded as functional words which do not carry any meaning.

Stemming: It is a process forfallingresultant words to their stem. The stemming program isgenerally referred to as stemmers or stemming algorithms. e.g. “established”, “establishment”, “establishing” is reduced to the stem “establish”.

Word Extension: This system enlarges the famous acronyms as well. The extensions are considered for Standard English words.

Redundant Words: If a word is being redundant in a tweet for more than several times repeatedly, occurrences of the word had been limited to several(i.e., two times) occurrences. E.g. "my mymymymy goodness" has been replaced by "my goodness". And a character like "sweeeeeet" has been replaced by "sweet".

C. Dynamic Natural Language Processing (DNLP)

The dynamic natural language processing method determining a sentence and distinguishing it from a phrase in a text is aproblematic task in natural language processing. Based on the dictionary, a sentence is a set of words that convey complete senses, has a main verb and starts with a capital letter. While a phrase is a small set of words within a sentence that makes a meaningful unit. This method applies Stanford Document Preprocessor that is a part of Stanford Parser, to each document in order to extract their sentences. After mining the sentences, DNLP uses a tokenizer to create a list of tokens. A tokenizer or lexer is a program or function that converts a sequence of characters into a sequence of tokens such that makes meaningful character string. In many languages including English, the whitespace and punctuation are used as aword or token delimiter and most of the tokenizers use them to split up the strings into tokens.

One specific application in DNLP is that can be used for the purpose is sentiment analysis. It can be used to identify and extract subjective information from the information source collected. With all these processes and methods, it is possible to build a system which can extract application dependent information, process it and produce the data which can be used for studying and deductions based on the information retrieved.

D. Cluster based Naive Bayesian Classification

The cluster based Naive Bayes classifier is an easy probabilistic classifier supported by Bayes' theorem with naive independence statements. In thisexpressions,it takes an assumption that a single word showing in a document doesn't change the probability of another word showing significance they are completely self-sufficient. To organize the work with naive Bayes there is arequirement for training data which are pre-classified. With the supportof the training data, the research work can make a decision to which cluster (positive or 1, negative or -1 and neutral or 0) for current document belongs to. To take a statement that with the stop words thrown absent of the document, the majority frequent expression in the document chooses to which class (1, 0 or -1) that document belongs to. Consequently, if a single name of the individual specific class shows with its synonyms. For every pre-classified class names, wordnet was used to create a synonym group for calculating the chances. There is a confident bound set where the outcomes under that direct to the classification ofthe document as neutral and not in any of the categories.

ALGORITHM:CLUSTER BASED NAIVE BAYESIAN CLASSIFICATION

Input: Given a set of online tweet review documents;

Output: Classification and Cluster result (positive, negative and neutral)

Step 1: For every preprocessed document in *N* cluster each one of them with its class type and find the amount of documents N_c in each class group.

Step 2: For every class,*c* calculate the chances of the class

Step 3: For every class and for each term in c_j, identify to which class belongs to using the following formula,

$$Class\ c = max\left[PC_j \sqcap P\left(\frac{x_i}{c_j}\right)\right]$$

Step 4: if c == 1

Cluster = positive tweet
Else if c == -1
Cluster = Negative tweet
Else if c==0
Cluster = Neutral
endif

Step 5: Arrange Cluster for the corresponding class

IV. SIMULATION RESULTS

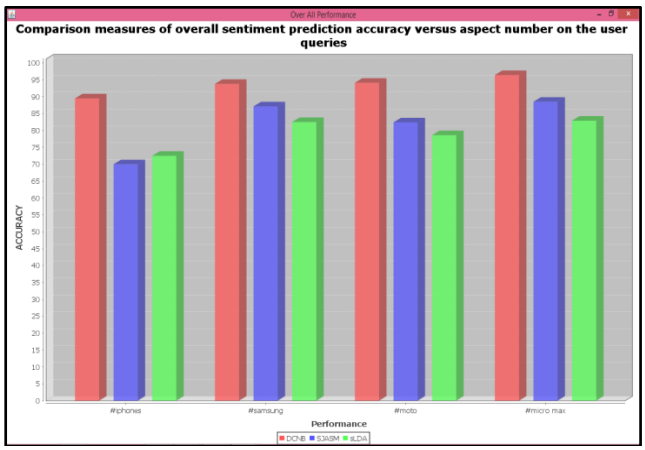
To the best of the information, there is no interpreted dataset for sentiment analysis retrieval in Twitter. Consequently, a new online review document for this job is created. Regarding 50 million document tweets are edged and guided by the Twitter API. This document a twitter search engine is performing to retrieve a user query for sentiment analysis. Estimation of overall sentiment prediction shows results on the user's input queries. DCNB outperforms all the benchmark models SJASM, sLDA across all number of test quires.

The computational accuracy of the classifier on the whole evaluation dataset, i.e.:

$$\text{Accuracy} = \frac{N(correct\ classifications)}{N(all\ classifications)}$$

Table 1: Comparison measures of overall sentiment prediction accuracy versus aspect number on the user queries.

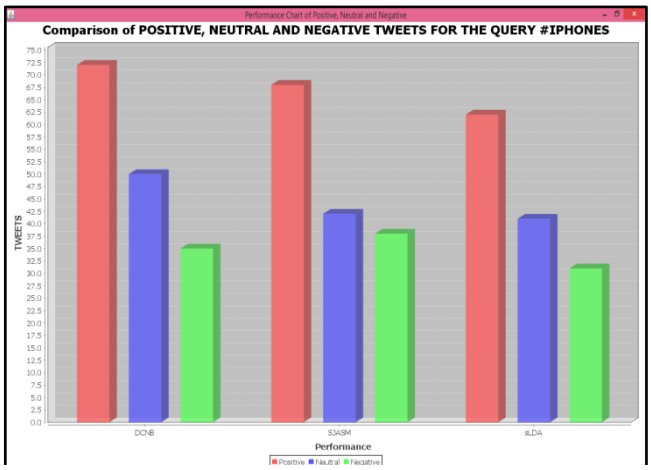
Methods	DCNB	SJASM	sLDA
#iphones	89	88	82
#samsung	93	87	84.5
#moto	94.13	86.8	83
#micromax	96.4	88.5	82.9



Comparison of overall sentiment prediction

Table 2: Comparison of Positive, Neutral and Negative Tweets for the Query #IPHONES

Methods	DCNB	SJASM	sLDA
Positive	**72**	**68**	62
Neutral	50	42	41
Negative	35	38	31



Comparison of Positive, Neutral and Negative measures

V. CONCLUSION

The main contribution of this research work is to implement unsupervised algorithm together with a Dynamic Cluster based Naive Bayesian Classification (DCNB) for sentiment analysis. Sentiment analysis (SA) is a framework of artificial intelligence and natural language processing that mechanically mines the sentiments

(positive, negative and neutral) from reviews. Sentiment Analysis is a very positive application for any associations that are looking for people's attitudes about their products and services. Since sentiment analysis naturally is a domain dependent approach that makes it problematic and expensive task, DCNB algorithm is used to train the Bayesian classifier on different domains. The results indicate that the DCNB is the most outperforming algorithm which increases the accuracy of classification by removing the redundant data and cluster'sthe tweets into positive, neutral and negative.

REFERENCES

[1] Hu. M and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '04), pp. 168-177, 2004.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in Proc. ACL-02 Conf. Empirical Methods Natural Language Process., 2002, pp. 79–86.

[3] Shen .D, J. Wu, B. Cao, J.-T. Sun, Q. Yang, Z. Chen, and Y. Li, "Exploiting Term Relationship to Boost Text Classification," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM '09), pp. 1637-1640, 2009.

[4] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in Proc. 4th Int. AAAI Conf. Weblogs Soc. Media, 2010, vol. 10, pp. 178–185.

[5] L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, "Predicting collective sentiment dynamics from time-series social media," in Proc. 1st Int. Workshop Issues Sentiment Discovery Opinion Mining, 2012, p. 6.

[6] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. 49th Annu. Meet. Assoc. Comput. Linguistics Human Language Technol., 2011, pp. 142–150.

[7] B. Yang and C. Cardie, "Context-aware learning for sentence-level sentiment analysis with posterior regularization," in Proc. 52nd Annu. Meet. Assoc. Comput. Linguistics, 2014, pp. 325–335.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[9] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 751–760.

[10] Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng, and Chunyan Miao, "Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach", IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 6, June 2017.

[11] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in Proc. 53rd Annu. Meet. Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Language Process. Jul. 2015, pp. 1014–1023.

[12] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," ACM Trans. Inf. Syst., vol. 26, no. 3, pp. 12:1–12:34, Jun. 2008.

[13] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proc. Conf. Human Language Technol. Empirical Methods Natural Language Process., 2005, pp. 347–354.

[14] A. Agarwal, F. Biadsy, and K. R. Mckeown, "Contextual phraselevel polarity analysis using lexical affect scoring and syntactic Ngrams," in Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics, 2009, pp. 24–32.

[15] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Clustering product features for opinion mining," in Proc. 4th ACM Int. Conf. Web Search Data Mining, 2011, pp. 347–354.